Tracing the origin of SARS-CoV-2: lessons learned from the past

Qihui Wang[1] , Hua Chen[2], Yi Shi[1], Alice C. Hughes[3], William J. Liu[4], Jingkun Jiang[5], George F. Gao[1], Yongbiao Xue[2], Yigang Tong[6]

[1] CAS Key Laboratory of Pathogen Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, 100101, China
[2] Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Centre for Bioinformation, Beijing, 100101, China
[3] Landscape Ecology Group, Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, 666303, China
[4] National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, 102206, China
[5] School of Environment, Tsinghua University, Beijing, 100084, China
[6] Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, P. R. China
Correspondence: Qihui Wang (wangqihui@im.ac.cn)

The origin of SARS-CoV-2 remains elusive. Understanding how, when, and where SARS-CoV-2 was transmitted from its natural reservoir to human beings is crucial for preventing future coronavirus outbreaks. With the lessons learned from the endless battle with pathogens and accumulated research data with regard to the origin and intermediate host, we present multiple potential locations as the natural reservoir of SARS-CoV-2.

Emerging and re-emerging infectious diseases pose a significant threat to human health, economy, and security worldwide. In recent years, we have witnessed the emergence of novel pathogens at an accelerating rate,[1] most of which are zoonotic pathogens, including Nipah virus, influenza virus, and especially, coronaviruses (CoVs).[2] After the outbreaks of severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), researchers worldwide have reached a consensus that the occurrence of the next CoV spillover event is only a matter of time, as supported by research data and the natural laws of pathogen emergence.[3] In other words, the outbreak of the SARS-CoV-2 is actually a gray rhino event that was predicted by professionals.

To change such an upward trend and prevent future spillover events, it is crucial to identify the origin and intermediate hosts of known pathogens. For this purpose, important lessons must be learned from the endless battle between humans and their pathogens.

First, determining the origins of a pathogen requires solid evidence. Specifically, a highly similar sequence-related virus must be identified from an animal that shares an ecological link with the virus' reservoir host or a known intermediate host. Here, we will use the origin tracing of MERS-CoV as an example. Strong evidence indicates that the 2012 MERS-CoV outbreak was driven by a dromedary-to-human spillover event,[4]

but the animal that transmitted MERS-CoV to dromedaries still remains unclear. Bats are hypothesized to be the natural reservoir for MERS-CoV because the bat CoV HKU4 displays sequence homology and similar receptor binding patterns with MERS-CoV. This suggests that MERS-CoV may be an HKU4-related virus that originated from bats. To date, however, no virus with a genome that is highly homologous to MERS-CoV has been identified from any bat species,[4] which prevents drawing a conclusion regarding to the origin of MERS-CoV. In contrast, another CoV, swine acute diarrhea syndrome coronavirus (SADS-CoV), which causes the death of piglets, was quickly determined as a bat-origin CoV after its outbreak because a highly similar virus (98.48% identity), bat CoV HKU2, was found in bats living in a cave near the infected pig farms.[4]

Second, tracing the origins of a virus could require decades of continuous research, but the accumulated data would form the foundation of future origin tracing capability. For example, it has long been known that the influenza A virus circulates in wild aquatic birds and can be transmitted to other avian and mammalian hosts by a number of processes, including mutation and reassortment.[5] In the past century, extensive surveillance of influenza A viruses in animals and humans has created an enormous amount of genome sequence data. Using the database that compiles these data, the origin of some newly emergent influenza A strain has been quickly traced, such as the H1N1 pandemic strain in 2009 and the H7N9 avian influenza strain in 2013.[6, 7]

Third, the location of the first outbreak might be far from the place of origin. Take human immunodeficiency virus (HIV) as an example. HIV was believed to have originated in the United States when it was first identified in the 1980s. Since then, scientists and health workers have become increasingly aware of HIV and officially recognized AIDS as a new human infectious disease. However, subsequent studies discovered a blood sample with HIV taken in 1959 from a man living in Kinshasa in the Democratic Republic of the Congo, which confirmed the first verified case of HIV in Africa.[8] Thus, the place where a new infectious disease is reported may not be the original place of disease occurrence.

To trace the origins of SARS-CoV-2, it is crucial to learn from history. First, the progenitor of the virus, which has strong similarity to SARS-CoV-2 must be found from a geographically and ecologically relevant animal before drawing conclusions. Second, origin tracing must not rush to a conclusion before accumulating sufficient evidence. Third, the fact that the location of the first outbreak might not be the place of origin must be kept in mind.

To find the progenitor of SARS-CoV-2 in animals, a number of SARS-related CoVs (sarbecoviruses) from around the world have been investigated, including RaTG13/RaTG15/RmYN02 (southern China), RshSTT182/RshSTT200 (Cambodia), Rc-o139 (Japan), RacCS203 (Thailand), BM48-31 (Bulgaria), and BtKY72 (Kenya).[9] Notably, all of these sarbecoviruses were discovered from bats of the *Rhinolophus* genus,[9] making *Rhinolophus* bats the potential reservoir hosts of SARS-CoV-2. However, as the closest known sarbecovirus related to SARS-CoV-2, RaTG13 still displays significant differences from SARS-CoV-2 with regard to its genome sequence, receptor binding pattern, and host range,[10] suggesting that bats as the potential natural

hosts of SARS-CoV-2 remains inconclusive. According to the World Health organization (WHO)-convened Global Study of Origins of SARS-CoV-2: China Part (hereafter referred to as the "WHO report"), direct zoonotic spillover is considered to be a possible-to-likely pathway.[9] Therefore, a global search for natural reservoirs with the potential to carry SARS-CoV-2-like viruses is urgently needed.

The WHO report also concluded that the introduction of SARS-CoV-2 through an intermediate host is considered to be a likely to very likely pathway.[9] To determine potential intermediate hosts of SARS-CoV-2, a number of mammalian species have been investigated, including domesticated animals (*e.g.*, horses, pigs, and cows), companion animals (*e.g.*, cats and dogs), and wild animals (*e.g.*, bats, pangolins, minks, foxes, and civets). Research data show that the angiotensin-converting enzyme 2 (ACE2) receptor from many of these species has an affinity for binding to the SARS-CoV-2 receptor-binding domain (RBD) similar to that of human ACE2, suggesting potential cross-species transmission paths between these animals to humans.[11] Among the possible intermediate hosts of SARS-CoV-2, pangolins and minks have attracted more attention than others. Pangolins have been found to host at least two CoVs, GX/P2V/2017 and GD/1/2019, that are closely related to SARS-CoV-2.[12] Alternately, minks might also be an intermediate host because the only reported SARS-CoV-2 outbreak in animals occurred in the mink population in Europe. This indicates that SARS-CoV-2 is well adapted to minks, and minks might have played an important role in the evolution of SARS-CoV-2.[13] All of these possibilities must be taken into consideration to unravel the mystery of the intermediate host of SARS-CoV-2.

The cross-species transmission of SARS-CoV-2 from the reservoir host to the intermediate host requires that the two hosts live in proximity and share ecological links. Considering the potential reservoir host and intermediate hosts, the location of origin of SARS-CoV-2 could be in regions where the distribution of *Rhinolophus* bats overlaps with that of pangolins, minks, or other potential intermediate hosts. Mustelids (which includes the mink) are distributed across the entire old-world. Therefore, we mapped the overall distribution area of 98 *Rhinolophus* species, eight pangolin species, and the wild European mink (*Mustela lutreola*), together with the main distribution area of mink farms.[13] We then marked the locations where bat sarbecoviruses were discovered and international flight routes to Wuhan (Figure 1). The distribution area of *Rhinolophus* species covers the southern portion of the Eurasian continent, the islands of Southeast Asia, and most of sub-Saharan Africa, which overlaps with that of pangolins in southern China, Southeast Asia, India, and sub-Saharan Africa. The European mink is distributed across Europe, which overlaps with the *Rhinolophus* distribution area in southern Europe. However, the majority of minks in Eurasia are the millions of American minks (*Neovison vison*) kept in mink farms in various of European countries and China,[13] which overlaps with the *Rhinolophus* distribution area in southern European countries such as Italy, Greece, Spain, and France, as well as some northern Chinese provinces.

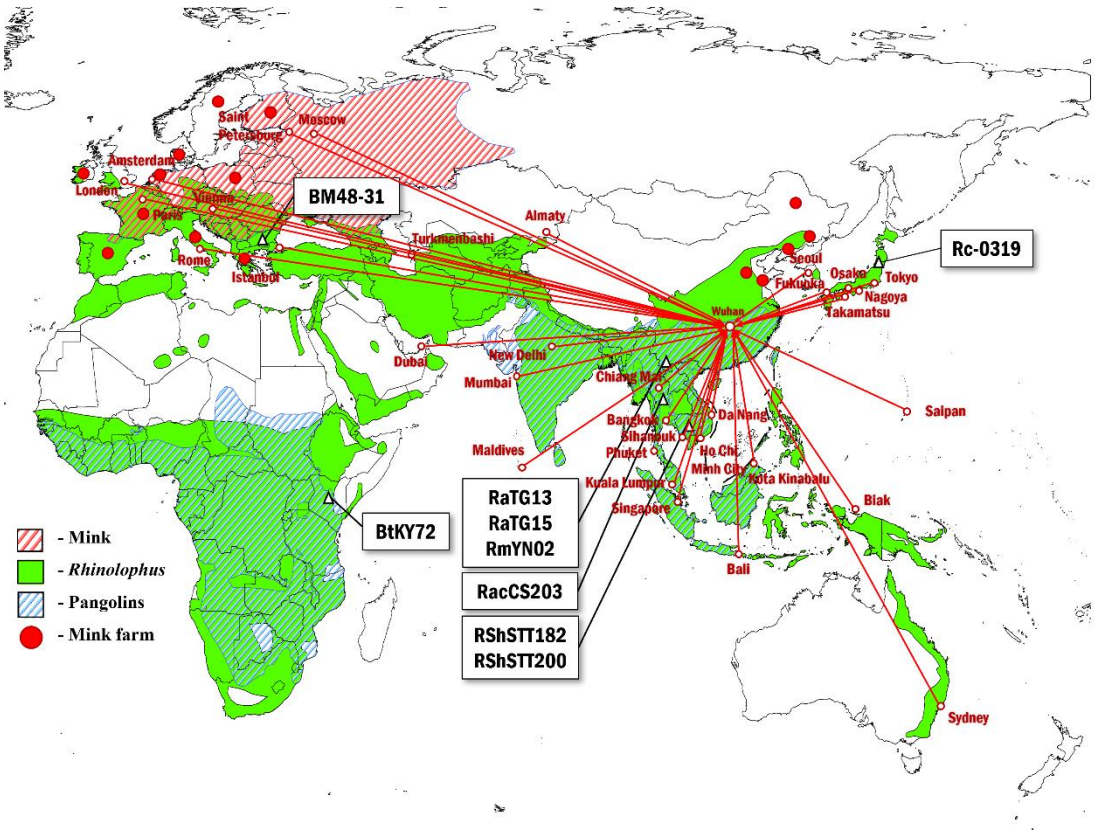

Figure 1: Distribution of *Rhinolophus*, pangolin and mink species, showing locations of bat sarbecoviruses discovered and the main distribution areas of mink farms.[13] Red lines indicate international flight routes to Wuhan. Animal distribution data are from the database of International Union for conservation of Nature (IUCN) Red list of Threatened Species (https://www.iucnredlist.org/). Air route information are from the website of Wuhan Tianhe Airport (http://www.whairport.com/).

These data suggest multiple locations worldwide where SARS-CoV-2 could be transmitted from its natural reservoir to intermediate hosts, before even considering other potential hosts and other intermediate hosts (such as other native carnivores), which are distributed across the old-world. Specifically, sarbecovirus spillover from *Rhinolophus* to pangolins could occur in Southeast Asia, southern China, India, and sub-Saharan Africa, while cross-species transmission from *Rhinolophus* to minks could occur in southern Europe. Both transmission routes could eventually lead to the adaptation of the viruses and potential human infection. Importantly, most of these regions show evidence of sarbecovirus circulation in bats, which allows multiple SARS-CoV-2-like viruses to evolve independently. Therefore, surveillance of sarbecoviruses needs to be conducted in *Rhinolophus* bats, pangolins, and minks from the abovementioned regions before determining the place of origin of SARS-CoV-2.

Aside from the distribution area of hosts, evolution analyses could also help to locate the place of origin of SARS-CoV-2. Specifically, accurate inference of the time to the most recent common ancestor (TMRCA) and initial evolutionary trajectories of the early SARS-CoV-2 sequences would facilitate unraveling the origin of SARS-CoV-2. The TMRCA of the early SARS-CoV-2 sequences was inferred to be November 28, 2019, with a 95% CI of [Oct 20, 2019, Dec 9, 2019], indicating that COVID-19 might

have originated from at an earlier time and outside of the Wuhan Seafood Market.[14] Furthermore, by constructing a haplotype network of the early SARS-CoV-2 genomes, the viral sequences can primarily be divided into two lineage clades, among which, the samples isolated from the Huanan Seafood Market mainly cluster with the descendant lineages rather than the ancestral lineages. This also indicates that the source of the CoV in the Market was likely imported from elsewhere.[15]

In addition, as a hub of international communication in central China, Wuhan received extensive international flights from cities around the world before the SARS-CoV-2 pandemic (Figure 1). Notably, many of these flights to Wuhan departed from Southeast Asian countries that overlap with the *Rhinolophus* and pangolin distributions, as well as multiple known sarbecoviruses. As mentioned in the WHO report, introduction through cold/food chain products is considered as a possible pathway. Therefore, before the pandemic, Wuhan was already at high risk of importing SARS-CoV-2 through cold chain cargo from other parts of the world.

In conclusion, as mentioned in the WHO report, it is possible-to-likely that SARS-CoV-2 was introduced by a direct zoonotic spillover, and it is likely to very likely that it was introduced through an intermediate host. Importantly, SARS-CoV-2 being introduced through cold/food chain products is possible, while a laboratory incident that led to the SARS-CoV-2 outbreak is extremely unlikely. More evidence needs to be collected to identify the origins, intermediate hosts, and transmission paths of SARS-CoV-2.[9] Tracing the origins and intermediate hosts of a virus is a difficult task. A solid conclusion is the result of an enormous amount of work, patience, global cooperation, some luck, and possibly decades of continuous research, as was accomplished for the influenza virus. Understanding the species ecology and the interaction between possible host species and the impacts of landscape management on future spillover risks are also important considerations for future research. Such work is indispensable for reducing the frequency of the inevitable pathogen emergences and the damage of outbreaks, for it is crucial to the common health of all mankind.

References

1 Gao GF. From "A"IV to "Z"IKV: attacks from emerging and re-emerging pathogens. *Cell* 2018; **172**:1157-1159.

2 Plowright RK, Parrish CR, McCallum H *et al.* Pathways to zoonotic spillover. *Nat Rev Microbiol* 2017; **15**:502-510.

3 Su S, Wong G, Shi W *et al.* Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol* 2016; **24**:490-502.

4 Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 2019; **17**:181-192.

5 Su S, Bi Y, Wong G, Gray GC, Gao GF, Li S. Epidemiology, evolution, and recent outbreaks of avian influenza virus in China. *J Virol* 2015; **89**:8671-8676.

6 Smith GJ, Vijaykrishna D, Bahl J *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 2009; **459**:1122-1125.

7 Gao R, Cao B, Hu Y *et al.* Human infection with a novel avian-origin influenza A (H7N9) virus. *N Engl J Med* 2013; **368**:1888-1897.
8 Sharp PM, Hahn BH. The evolution of HIV-1 and the origin of AIDS. *Philos Trans R Soc Lond B Biol Sci* 2010; **365**:2487-2494.
9 WHO-convened global study of origins of SARS-CoV-2: China Part. Available from: https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.
10 Liu K, Pan X, Li L *et al.* Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species. *Cell* 2021; **184**:3438-3451 e3410.
11 Wu L, Chen Q, Liu K *et al.* Broad host range of SARS-CoV-2 and the molecular basis for SARS-CoV-2 binding to cat ACE2. *Cell Discov* 2020; **6**:68.
12 Lam TT, Jia N, Zhang YW *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* 2020; **583**:282-285.
13 Fenollar F, Mediannikov O, Maurin M *et al.* Mink, SARS-CoV-2, and the human-animal interface. *Front Microbiol* 2021; **12**:663815.
14 Liu Q, Zhao S, Shi CM *et al.* Population genetics of SARS-CoV-2: disentangling effects of sampling bias and infection clusters. *Genomics Proteomics Bioinformatics* 2020.
15 Yu WB, Tang GD, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zool Res* 2020; **41**:247-257.